011404945 **Image available**
WPI Acc No: 1997-382852/ 199735
XRPX Acc No: N97-318669

Space detection method for documents containing both English and Japanese characters - by evaluating length of extracted object range to determine if its length is constant or proportional

Patent Assignee: RICOH KK (RICO)
Number of Countries: 001 Number of Patents: 001
Patent Family:

Patent No	Kind	Date	Applicat No	Kind	Date	Week	
JP 9167206	A	19970624	JP 95328737	A	19951218	199735	B

Priority Applications (No Type Date): JP 95328737 A 19951218
Patent Details:

Patent No	Kind	Lan	Pg	Main IPC	Filing Notes
JP 9167206	A		13		

Abstract (Basic): JP 9167206 A

The method involves extracting an object range consisting of several English character rows. The length of the extracted object range is evaluated when performing a length format evaluation process (301) to determine if it is a proportional length or a constant length.

When the length of the object range is constant, a constant length character string detection process (303) detects a space for every object range. When the length of the object range is proportional, a proportional length character string detection process (302) detects a space for every object range.

ADVANTAGE - Provides accurate space detection by using length of extracted character string.

Dwg.1/9

Title Terms: SPACE; DETECT; METHOD; DOCUMENT; CONTAIN; ENGLISH; JAPAN; CHARACTER; EVALUATE; LENGTH; EXTRACT; OBJECT; RANGE; DETERMINE; LENGTH; CONSTANT; PROPORTION
Derwent Class: T01; T04
International Patent Class (Main): G06K-009/36
International Patent Class (Additional): G06K-009/62
File Segment: EPI
Manual Codes (EPI/S-X): T01-J10B2; T04-D03; T04-D04

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平9－167206

(43)公開日　平成９年(1997)６月24日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｋ	9/36			Ｇ０６Ｋ	9/36	
	9/62	６１０	9061－5H		9/62	６１０Ｂ

審査請求　未請求　請求項の数７　ＯＬ　（全 13 頁）

(21)出願番号　特願平7－328737

(22)出願日　平成７年(1995)12月18日

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号

(72)発明者　嶺脇　隆邦
東京都大田区中馬込１丁目３番６号　株式会社リコー内

(74)代理人　弁理士　鈴木　誠　（外１名）

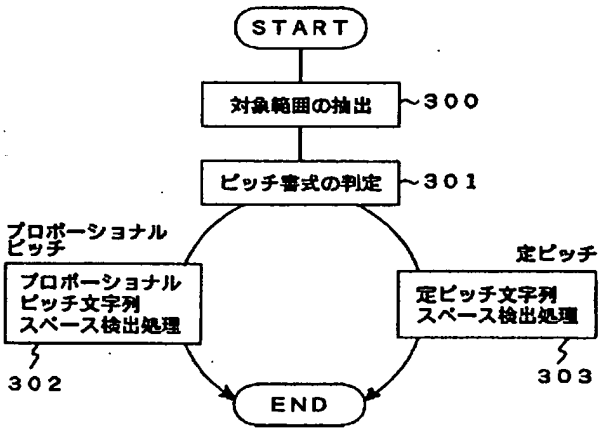
(54)【発明の名称】　日英混在文書のスペース検出方法、ピッチ書式判定方法、定ピッチ英数文字列のスペース検出方法、及びプロポーショナルピッチ英数文字列のスペース検出方法

(57)【要約】

【課題】　日本文字列と混在した英数文字列中のスペースを高精度に検出する。

【解決手段】　文字認識装置において、認識結果から英数文字列を抽出し（３００）、英数文字列毎にピッチ書式を判定する（３０１）。各英数文字列に対し、そのピッチ書式に応じプロポーショナルピッチ用スペース検出処理（３０２）又は定ピッチ用スペース検出処理（３０３）を適用し、高精度にスペースを検出する。

スペース検出処理フロー



【特許請求の範囲】
【請求項１】　文書の画像より文字を切り出し認識する文字認識装置において、英日混在の文書の文字認識結果に基づいて英数文字列を対象範囲として抽出する対象範囲抽出処理と、該対象範囲抽出処理により抽出された各対象範囲毎にピッチ書式が定ピッチかプロポーショナルピッチかを判定するピッチ書式判定処理と、該ピッチ書式判定処理により定ピッチと判定された各対象範囲毎にスペースを検出する定ピッチ文字列スペース検出処理と、該ピッチ書式判定処理によりプロポーショナルピッチと判定された各対象範囲毎にスペースを検出するプロポーショナルピッチ文字列スペース検出処理とを有することを特徴とする日英混在文書のスペース検出方法。
【請求項２】　文書の画像より文字を切り出し認識する文字認識装置において、英日混在の文書の文字認識結果に基づいて英数文字列を対象範囲として抽出する対象範囲抽出処理と、該対象範囲抽出処理により抽出された各対象範囲毎にピッチ書式が定ピッチかプロポーショナルピッチかを判定するピッチ書式判定処理と、該ピッチ書式判定処理により定ピッチと判定された各対象範囲毎にスペースを検出する定ピッチ文字列スペース検出処理と、該ピッチ書式判定処理によりプロポーショナルピッチと判定された各対象範囲毎にスペースを検出するプロポーショナルピッチ文字列スペース検出処理とを有し、該ピッチ書式判定処理において、各対象範囲に対し、各対象範囲毎に切り出し情報に基づき決定される標準文字サイズに対する文字矩形間隔の比である文字矩形間隔比が所定の閾値より小さい文字矩形間隔のほうが、文字矩形間隔比が該閾値以上の文字矩形間隔より多数であるならばプロポーショナルピッチと判定し、そうでなければ定ピッチと判定することを特徴とする日英混在文書のスペース検出方法。
【請求項３】　文書の画像より文字を切り出し認識する文字認識装置において、英日混在の文書の文字認識結果に基づいて英数文字列を対象範囲として抽出する対象範囲抽出処理と、該対象範囲抽出処理により抽出された各対象範囲毎にピッチ書式が定ピッチかプロポーショナルピッチかを判定するピッチ書式判定処理と、該ピッチ書式判定処理により定ピッチと判定された各対象範囲毎にスペースを検出する定ピッチ文字列スペース検出処理と、該ピッチ書式判定処理によりプロポーショナルピッチと判定された各対象範囲毎にスペースを検出するプロポーショナルピッチ文字列スペース検出処理とを有し、該定ピッチ文字列スペース検出処理において、注目した文字間の前後の文字間の文字矩形ピッチのうちの小さい方の文字矩形ピッチを注目した文字間の基準ピッチとし、注目した文字間の文字矩形ピッチの基準ピッチとの比が所定の閾値より大きいときに、注目した文字間にスペースが存在すると判定することを特徴とする日英混在文書のスペース検出方法。
【請求項４】　文書の画像より文字を切り出し認識する文字認識装置において、英日混在の文書の文字認識結果に基づいて英数文字列を対象範囲として抽出する対象範囲抽出処理と、該対象範囲抽出処理により抽出された各対象範囲毎にピッチ書式が定ピッチかプロポーショナルピッチかを判定するピッチ書式判定処理と、該ピッチ書式判定処理により定ピッチと判定された各対象範囲毎にスペースを検出する定ピッチ文字列スペース検出処理と、該ピッチ書式判定処理によりプロポーショナルピッチと判定された各対象範囲毎にスペースを検出するプロポーショナルピッチ文字列スペース検出処理とを有し、該プロポーショナルピッチ文字列スペース検出処理において、各対象範囲毎に切り出し情報に基づき決定される標準文字サイズに対する文字矩形間隔の比である文字矩形間隔比が、所定の閾値より大きいときに、対応する文字間にスペースが存在すると判定することを特徴とする日英混在文書のスペース検出方法。
【請求項５】　文書の画像より文字を切り出し認識する文字認識装置において、認識された各英数文字列に対し、各英数文字列毎に切り出し情報に基づいて決定される標準文字サイズに対する文字矩形間隔の比である文字矩形間隔比が所定の閾値より小さい文字矩形間隔のほうが、文字矩形間隔比が該閾値以上の文字矩形間隔より多数であるならばプロポーショナルピッチと判定し、そうでなければ定ピッチと判定することを特徴とするピッチ書式判定方法。
【請求項６】　文書の画像より文字を切り出し認識する文字認識装置において、認識された定ピッチの各英数文字列において、注目した文字間の前後の文字間の文字矩形ピッチのうちの小さい方の文字矩形ピッチを注目した文字間の基準ピッチとし、注目した文字間の文字矩形ピッチの基準ピッチとの比が所定の閾値より大きいときに、注目した文字間にスペースが存在すると判定することを特徴とする定ピッチ英数文字列のスペース検出方法。
【請求項７】　文書の画像より文字を切り出し認識する文字認識装置において、認識されたプロポーショナルピッチの各英数文字列において、各英数文字列毎に切り出し情報に基づき決定される標準文字サイズに対する文字矩形間隔の比である文字矩形間隔比が、所定の閾値より大きいときに、対応する文字間にスペースが存在すると判定することを特徴とするプロポーショナルピッチ英数文字列のスペース検出方法。
【発明の詳細な説明】
【０００１】
【発明の属する分野】本発明は、文字認識装置におけるスペース検出処理に係り、特に、日英混在の文書に対するスペース検出処理に関する。
【０００２】
【従来の技術】文書を文字認識によってテキストデータ

に変換し、このテキストデータから文書を再現できるようにする場合等には、文字認識に際し、文書中のスペースも検出する必要がある。このような文書中のスペース検出に関する従来技術としては次に述べるようなものが知られている。

【０００３】従来技術１：文字間の白画素数を計数し、計数値を標準文字ピッチで割ることによりスペースコード数を決定する（特願昭６３－１４２８２号）。

【０００４】従来技術２：文書のピッチ書式（定ピッチ／プロポーショナルピッチ）を判定する。そして、判定したピッチ書式に応じた方法でスペースを検出する（特願平１－１６１１７６号）。

【０００５】従来技術３：日英文字混在の日本語文章において、文字間空白幅と文字幅の情報を比較しスペースの種類と有無を判定するが、スペースの前後にある文字の種類を参照してスペースの有効／無効を判定する（特願平２－２１４１３６号）。

【０００６】従来技術４：日英文字混在の日本語文章において、文字間空白幅と文字の標準サイズを比較し、スペースの種類と有無を判定するが、文字間空白に隣接する文字に応じて文字間空白幅を補正し、この補正後の値を標準文字サイズと比較する（特願平３－１８４７６号）。

【０００７】

【発明が解決しようとする課題】日本文字と英数字（英文字と数字）が混在する文書では、日本文字と英数字のフォントの違い、全角文字と半角文字という文字サイズの違い、定ピッチとプロポーショナルピッチというピッチ書式の違いが存在し、しかも、このような違いが文書中の不特定の部分に生じる。そして、スペースの幅も場所によって違いが生じる。このことは図９に示す印字サンプルをみれば容易に理解されよう。図９において、全て全角文字で定ピッチ印字されたサンプル１では、「Ｒｉｃｏｈ」の「ｈ」と次の「Ｐ」との間にだけスペースがある。これに対して英数字がプロポーショナルピッチで印字されたサンプル２では、「新型」の「型」と次の「Ｒ」との間、及び「Ｒｉｃｏｈ」の「ｈ」と次の「Ｐ」にスペースがあり、しかも、そのスペースはサンプル１のスペースより間隔が狭い。

【０００８】このような日英混在文書に対しては、従来技術では高精度のスペース検出が困難な場合があった。また、行中で日本文字と英数字が混在するような場合にピッチ書式を的確に判定する方法は知られていなかった。従来技術２は、文書全体又は行全体が英文であると仮定し、英単語単位の処理となっているため、行中に部分的に現れる英単語中のスペースを精度よく検出できなかった。定ピッチ文字列中のスペース検出に関しては、定ピッチ文字列の文字間隔の変動が大きいため、全角サイズのスペースは検出できても、文字間隔を補正する従来技術４によっても半角スペースの検出精度が上がらなかった。プロポーショナルピッチ文字列中のスペース検出に関しては、プロポーショナルピッチ文字列の文字間隔が狭いので、従来技術４によってもスペースを検出できなかった。

【０００９】本発明の目的は、日本文字と英数字が行中に混在するような日英混在文書に対し高精度のスペース検出が可能な改良した方法、並びに、このスペース検出方法のために好適なピッチ書式の判定方法、定ピッチ文字列中のスペースの検出方法、及びプロポーショナルピッチ文字列中のスペースの検出方法を提供することにある。

【００１０】

【課題を解決するための手段】請求項１記載の発明による日英混在文書のスペース検出方法は、文書の画像より文字を切り出し認識する文字認識装置において、英日混在の文書の文字認識結果に基づいて英数文字列を対象範囲として抽出する対象範囲抽出処理と、該対象範囲抽出処理により抽出された各対象範囲毎にピッチ書式が定ピッチかプロポーショナルピッチかを判定するピッチ書式判定処理と、該ピッチ書式判定処理により定ピッチと判定された各対象範囲毎にスペースを検出する定ピッチ文字列スペース検出処理と、該ピッチ書式判定処理によりプロポーショナルピッチと判定された各対象範囲毎にスペースを検出するプロポーショナルピッチ文字列スペース検出処理とを有することを特徴とするものである。

【００１１】請求項２記載の発明は、請求項１記載の発明の日英混在文書のスペース検出方法において、ピッチ書式判定処理で、各対象範囲に対し、各対象範囲毎に切り出し情報に基づき決定される標準文字サイズに対する文字矩形間隔の比である文字矩形間隔比が所定の閾値より小さい文字矩形間隔のほうが、文字矩形間隔比が該閾値以上の文字矩形間隔より多数であるならばプロポーショナルピッチと判定し、そうでなければ定ピッチと判定することを特徴とするものである。

【００１２】請求項３記載の発明は、請求項１記載の発明の日英混在文書のスペース検出方法において、定ピッチ文字列スペース検出処理で、注目した文字間の前後の文字間の文字矩形ピッチのうちの小さい方の文字矩形ピッチを注目した文字間の基準ピッチとし、注目した文字間の文字矩形ピッチの基準ピッチとの比が所定の閾値より大きいときに、注目した文字間にスペースが存在すると判定することを特徴とするものである。

【００１３】請求項４記載の発明は、請求項１記載の発明の日英混在文書のスペース検出方法において、プロポーショナルピッチ文字列スペース検出処理で、各対象範囲毎に切り出し情報に基づき決定される標準文字サイズに対する文字矩形間隔の比である文字矩形間隔比が、所定の閾値より大きいときに、対応する文字間にスペースが存在すると判定することを特徴とするものである。

【００１４】請求項５記載の発明のピッチ書式判定方法

は、文書の画像より文字を切り出し認識する文字認識装置において、認識された各英数文字列に対し、各英数文字列毎に切り出し情報に基づいて決定される標準文字サイズに対する文字矩形間隔の比である文字矩形間隔比が所定の閾値より小さい文字矩形間隔のほうが、文字矩形間隔比が該閾値以上の文字矩形間隔より多数であるならばプロポーショナルピッチと判定し、そうでなければ定ピッチと判定することを特徴とする。

【００１５】請求項６記載の発明の定ピッチ英数文字列のスペース検出方法は、文書の画像より文字を切り出し認識する文字認識装置において、認識された定ピッチの各英数文字列において、注目した文字間の前後の文字間の文字矩形ピッチのうちの小さい方の文字矩形ピッチを注目した文字間の基準ピッチとし、注目した文字間の文字矩形ピッチの基準ピッチとの比が所定の閾値より大きいときに、注目した文字間にスペースが存在すると判定することを特徴とするものである。

【００１６】請求項７記載の発明のプロポーショナルピッチ英数文字列のスペース検出方法は、文書の画像より文字を切り出し認識する文字認識装置において、認識されたプロポーショナルピッチの各英数文字列において、各英数文字列毎に切り出し情報に基づき決定される標準文字サイズに対する文字矩形間隔の比である文字矩形間隔比が、所定の閾値より大きいときに、対応する文字間にスペースが存在すると判定することを特徴とするものである。

【００１７】

【発明の実施の形態】本発明の実施の形態を明らかにするため、図面を用いて本発明の一実施例を説明する。図１は本発明の一実施例の全体的処理フローを示し、図２は本発明の一実施例のための装置構成例を示す。図１中のスペース検出処理のフローを図３に示す。図３中のピッチ書式判定処理のフローを図４に、プロポーショナルピッチ文字列スペース検出処理のフローを図５に、定ピッチ文字列スペース検出処理のフローを図６に、それぞれ示す。図７及び図８はスペース検出処理の説明のための図である。

【００１８】初めに図１及び図２を参照し、全体的処理フローを装置構成と関連付けて説明する。まず、画像入力部２００により文書画像のデータを入力し、これをバス２１３を経由して画像メモリ２０８に格納する（ステップ１００）。画像が入力されると、行・文字切り出し部２０２において、画像メモリ２０８内の文書画像中の文字行と文字を切り出し、切り出した文字行及び文字の領域の情報を切り出しメモリ２０９に格納する（ステップ１１０）。ここに格納される切り出し情報は、例えば、文字行の始点と終点の座標あるいは始点座標と幅、文字の外接矩形（文字矩形）の対角頂点の座標やサイズ情報である。これらの情報は、各文字矩形と、それが所属する文字行との対応関係が識別できるような形で作成され格納されることは当然である。そして、文字認識部２０３において、切り出しメモリ２０９内の切り出し情報を参照し、画像メモリ２０８より文字画像を取り込み、それを文字辞書メモリ２１０内の文字辞書と比較することにより文字画像に対する文字コードを決定し、それを結果メモリ２１１に格納する（ステップ１２０）。なお、文字認識部２０３において、単語や文法等の言語知識を利用して、文字辞書との比較による認識結果に対する修正等の後処理を行ってもよい。

【００１９】文字認識部２０３による認識処理が文書画像の全体あるいは一部について終了した段階で、スペース検出処理（ステップ１３０）が開始する。このスペース検出処理（ステップ１３０）は、対象範囲抽出部２０４、ピッチ書式判定部２０５、定ピッチ文字列スペース検出部２０６及びプロポーショナルピッチ文字列スペース検出部２０７により実行され、その際に切り出しメモリ２０９、結果メモリ２１１及びワークメモリ２１２が参照される。スペース検出結果は結果メモリ２１１に格納される。スペース検出処理が終了すると、結果メモリ２１１内のデータが外部に出力され（ステップ１４０）、処理全体が終了する。

【００２０】スペース検出処理（ステップ１３０）の概略は図３に示すとおりである。まず、対象範囲抽出部２０４において、結果メモリ２１１内の文字認識結果（文字コード）を参照し、スペース検出処理の対象範囲としての英数文字列を抽出する（ステップ３０１）。抽出した対象範囲に関する情報はワークメモリ２１２に保存される。次に、ピッチ書式判定部２０５において、対象範囲の文字に関して切り出しメモリ２０９及び結果メモリ２１１の内容を参照し、対象範囲がプロポーショナルピッチであるか、定ピッチであるかを判定する（ステップ３０１）。判定結果はワークメモリ２１２に保存される。プロポーショナルピッチと判定された対象範囲に対しては、プロポーショナルピッチ文字列スペース検出部２０７において切り出しメモリ２０９及び結果メモリ２１１の内容を参照して、プロポーショナルピッチ用の方法によりスペース検出を行い、検出したスペースをスペースコードとして結果メモリ２１１に書き込む（ステップ３０２）。定ピッチと判定された対象範囲に対しては、定ピッチ文字列スペース検出部２０６において切り出しメモリ２０９及び結果メモリ２１１の内容を参照して、定ピッチ用の方法によりスペース検出を行い、検出したスペースをスペースコードとして結果メモリ２１１に書き込む（ステップ３０３）。ここに述べた処理の具体的な内容について、図４乃至図８を参照し説明する。

【００２１】図３の（ａ）に示すような文字行を処理する場合を考える。この文字行には、スペースを含む部分が２カ所ある。一つはプロポーショナルピッチ（フォントはTimes－Roman）で印字された「Ｐｒｅｔｅｒ　５５」の「Ｐｒｅｔｅｒ」と「５５」の間である。もう一

つは、定ピッチ（フォントはCourier）で印字された「Ｉｍａｇｉｏ　７７」の「Ｉｍａｇｉｏ」と「７７」の間である。この文字行に対する文字認識（ステップ１２０）の結果はスペースを含まず、「新型コピー、Ｐｒｅｔｅｒ５５とＩｍａｇｉｏ７７発売」となる。

【００２２】対象範囲抽出処理（ステップ３００）では、そのようなスペースを含まない文字認識結果を参照し、英数字（英字又は数字）が５文字以上連続している文字列を対象範囲として抽出する。したがって、図７の（ａ）に示す文字行では、図７の（ｂ）に示すように「Ｐｒｅｔｅｒ５５」と「Ｉｍａｇｉｏ７７」の二つの文字列が対象範囲として抽出される。なお、抽出判定のための文字長は５文字に限定されるものではなく、処理する文書の種類に応じて適宜変更してよい。また、対象範囲の決定方法そのものも適宜変更し得るものである。

【００２３】このようにして対象範囲として抽出された英数文字列について、ピッチ書式判定（ステップ３０１）を行う。図４はそのフローチャートである。

【００２４】まず、対象範囲に関する標準文字サイズを決定する（ステップ４００）。具体的には、対象範囲内における最大の文字矩形高さの１．２５倍（文字サイズＡ）と、対象範囲が含まれる文字行内の最大の文字矩形高さ（文字サイズＢ）を求める。文字サイズＡ，Ｂの大きい方の値を標準文字サイズとする。ただし、対象範囲内に「ｊ」の文字矩形がある場合、その文字矩形は標準文字サイズの決定には利用しない。図８は、このような標準文字サイズ決定の説明図である。

【００２５】図７に示した一つ目の対象範囲では、「Ｐ」の文字矩形の高さが最大であるので、その高さを１．２５倍した値と、文字行内で最大の「新」の文字矩形高さとを比較し、大きい方の値が標準文字サイズとされる。なお、他の方法によって標準文字サイズを決定してもよい。

【００２６】以上のようにして対象範囲に対する標準文字サイズを決定すると、次に、この標準文字サイズを用いて、対象範囲の先頭より、隣接する文字矩形のペアを選び、その文字矩形間の空白部の幅、すなわち文字矩形間隔（図７の（ｃ）参照）の標準文字サイズに対する比（文字矩形間隔比＝文字矩形間隔／標準文字サイズ）を計算する（ステップ４０５，４１０）。計算した文字矩形間隔比は、ワークメモリ２１２に保存される。文字矩形間隔は、切り出しメモリ２０９に保存されている二つの文字矩形の座標から算出される。ただし、切り出し処理の段階で文字矩形間隔を求めておいてもよく、その場合は、切り出しメモリ２０９より読み出した文字矩形間隔を文字矩形間隔比の計算に用いればよい。次に、このようにして計算した文字矩形間隔比を所定の閾値ＴＨ１と比較する（ステップ４１５）。この閾値ＴＨ１は例えば０．１７である（この値に限らないが、一般的文書では、このあたりの値が適当であることが実験により確認された。ただし、処理する文書に応じて調整するのが好ましい）。文字矩形間隔比が閾値ＴＨ１より小さいときにカウンタＮＳ（初期値は０）を１だけインクリメントし（ステップ４２０）、文字矩形間隔比が閾値ＴＨ１以上のときにカウンタＮＬ（初期値は０）を１だけインクリメントする（ステップ４２１）。次の文字矩形ペアが残っている場合は、ステップ４０５からステップ４２０の処理ループを繰り返す。

【００２７】最後の文字矩形ペアまで処理が終了すると、ステップ４０５からステップ４２０又は４２１までの処理ループを抜けステップ４２５に進む。このステップ４２５では、カウンタＮＳの値とカウンタＮＬの値を比較する（ステップ４２５）。カウンタＮＳの値は対象範囲内における文字矩形間隔比が閾値ＴＨ１より小さい、つまり「標準文字サイズ＊ＴＨ１」より狭い文字矩形間隔の個数であり、カウンタＮＬの値は広い文字矩形間隔の個数である。ゆえに、ステップ４２５では、狭い文字矩形間隔が過半数であるか否かを調べている。ＮＳ＞ＮＬつまり狭い文字矩形間隔が過半数と判定した場合には対象範囲の文字列のピッチ書式をプロポーショナルピッチに設定し、そのフラグ情報をワークメモリ２１２に対象範囲と対応付けて書き込み（ステップ４３０）、そうでない場合、つまり広い文字矩形間隔が半数以下である場合には、ピッチ書式を定ピッチに設定し、そのフラグ情報をワークメモリ２１２に対象範囲と対応付けて書き込む（ステップ４３５）。このフラグ情報にしたがって、次の処理として、プロポーショナルピッチ文字列スペース検出処理（ステップ３０２）又は定ピッチ文字列スペース検出処理（ステップ３０３）が選択される。

【００２８】図７の（ａ）に示した例では、最初の対象範囲「Ｐｒｅｔｅｒ５５」は図７（ｃ）に示すようにプロポーショナルピッチと判定され、もう一つの対象範囲「Ｉｍａｇｉｏ７７」は定ピッチと判定される。

【００２９】プロポーショナルピッチと判定された対象範囲に対するスペース検出処理（ステップ３０２）の処理内容について説明する。図５は、そのフローチャートである。

【００３０】ワークメモリ２１２に保存されている対象範囲の文字矩形間隔比を取り出し、それを所定の閾値ＴＨ２と比較する（ステップ５００，５０５）。この閾値ＴＨ２は例えば０．２９である（この値に限らないが、一般的文書では、このあたりの値が適当であることが実験により確認された。ただし、処理する文書に応じて調整するのが好ましい）。文字矩形間隔比がＴＨ２以下であれば、つまり文字矩形間隔が「標準文字サイズ＊ＴＨ２」より狭いときには、対応した文字矩形ペアの間にスペースは無いと判断される。文字矩形間隔比がＴＨ２より大きいときには、つまり文字矩形間隔が「標準文字サイズ＊ＴＨ２」より広いときには、対応した文字矩形ペアの間にスペースが存在すると判断され、対応する認識

結果文字コードの間にスペースコードが挿入される（ステップ５１０）。同様の処理が対象範囲内の全ての文字矩形間隔比について実行される。

【００３１】図７の（ａ）に示した例では、図７の（ｄ）に示すように、プロポーショナルピッチと判定された対象範囲において「Ｐｒｅｔｅｒ」の末尾の「ｒ」と次の「５」との間の文字矩形間隔比だけがＴＨ２を超えるので、その位置にスペースコードが挿入される。

【００３２】定ピッチと判定された対象範囲に対するスペース検出処理（ステップ３０３）の処理内容について説明する。図６は、そのフローチャートである。

【００３３】対象範囲内のすべての文字間の文字矩形ピッチを計算し、結果をワークメモリ２１２に保存する（ステップ６００）。すなわち、文字間を介して隣接する文字矩形の中心を計算し、その中心の間隔を当該文字間の文字矩形ピッチとして求める。この計算には、切り出しメモリ２０９に保存されている文字矩形の座標データを用いる。ただし、切り出し処理段階で予め隣接した文字矩形の中心間隔を計算して切り出しメモリ２０９に保存しておき、その値を読み出すようにしてもよい。

【００３４】このような文字矩形ピッチの計算を終わると、対象範囲の先頭より、注目する一つの文字間を選び（ステップ６０５）、注目文字間の前の文字間の文字矩形ピッチと、注目文字間の後の文字間の文字矩形ピッチとを比較し、小さい方の文字矩形ピッチを注目文字間に対する基準ピッチとする（ステップ６１０）。注目文字間が対象範囲の最初の文字間であるときには、注目文字間の前には文字間が存在しないので、注目文字間の後の文字間の文字矩形ピッチを基準ピッチとする。同様に、注目文字間が対象範囲の最後の文字間であるときには、その前の文字間の文字矩形ピッチを基準ピッチとする。

【００３５】次に、注目文字間の文字矩形ピッチを基準ピッチで割った値を文字矩形ピッチ比（＝文字矩形ピッチ／基準ピッチ）として計算する（ステップ６１５）。そして、文字矩形ピッチ比が所定の閾値ＴＨ３より大きいか、すなわち注目文字間の文字矩形ピッチが「基準ピッチ＊ＴＨ３」より広いか判定する（ステップ６２０）。閾値ＴＨ３は例えば１．８である（この値に限らないが、一般的文書では、このあたりの値が、全角文字列中の全角スペースも半角文字列中の半角スペースも検出可能であり、適当であることが実験により確認された。ただし、処理する文書に応じて調整するのが好ましい）。

【００３６】文字矩形ピッチ比がＴＨ３以下であれば、注目文字間にスペースが存在しないと判断される。文字矩形ピッチ比がＴＨ３より大きいときには、注目文字間にスペースが存在すると判断されるので、結果メモリ２１１内の対応した文字コードの間にスペースコードが挿入される（ステップ６２５）。同様の処理が対象範囲の最後の文字間まで繰り返される。

【００３７】図７の（ａ）に示した例では、図７の（ｄ）に示すように、定ピッチと判定された対象範囲において「Ｉｍａｇｉｏ」の最後の「ｏ」と次の「７」の文字間だけ文字矩形ピッチ比がＴＨ３を超えるので、その文字間にスペースが有ると判断されスペースコードが挿入される。

【００３８】

【発明の効果】請求項１記載の発明によれば、文字認識結果より英数文字列を対象範囲として抽出し、対象範囲毎にピッチ書式を判別し、各対象範囲に対して、そのピッチ書式別のスペース検出処理を行うため、行中に英数文字列と日本文字列とが混在し、しかも定ピッチの英数文字列とプロポーショナルピッチの英数文字列が混在するような日英混在文書に対しても、精度のよいスペース検出が可能になる。

【００３９】請求項２記載の発明によれば、対象範囲として抽出した英数文字列毎に標準文字サイズを決定し、英数文字列における文字矩形間隔の標準文字サイズに対する相対的な広狭の割合に着目してピッチ書式を判定するため、英数文字列が日本文字列と混在し、しかも全角文字と半角文字の英数文字列が混在するような場合にも、英数文字列のピッチ書式が定ピッチであるかプロポーショナルピッチであるかを精度よく判別でき、したがって日英混在文書中の英数文字列に対し、そのピッチ書式にあったスペース検出処理を的確に適用できるため、定ピッチとプロポーショナルピッチの英数文字列が混在した文書中のスペースを高精度に検出することができる。

【００４０】定ピッチ英数文字列の文字矩形ピッチは文字によって大きく変動するため、この変動が反映されない一定の基準ピッチと文字矩形ピッチとの相対的な広狭を調べても、スペースを精度よく検出できない。これに対し、請求項３記載の発明によれば、注目した文字間の前と後の文字間の文字矩形ピッチの中の小さい方を、注目した文字間に対する基準ピッチとして用いるため、文字矩形ピッチの変動が大きい定ピッチ英数字文字列中のスペースを精度よく検出することができ、したがって定ピッチの英数文字列が混在した文書に対するスペース検出精度を上げることができる。

【００４１】プロポーショナルピッチ英数文字列は文字矩形間隔が狭いため、ピッチ書式や文字種を考慮しない標準文字サイズと文字矩形間隔とを比較する方法では、スペース検出が難しかった。請求項４記載の発明によれば、個々のプロポーショナルピッチ英数文字列毎に標準文字サイズを決定し、標準文字サイズに対する文字矩形間隔の比を閾値処理するため、プロポーショナルピッチの英数文字列中のスペースを精度よく検出でき、したがってプロポーショナルピッチの英数文字列が混在した文書のスペース検出精度を上げることができる。

【００４２】請求項５記載の発明によれば、英数文字列

毎に標準文字サイズを決定し、英数文字列における文字矩形間隔の標準文字サイズに対する相対的な広狭の割合に着目してピッチ書式を判定するため、英数文字列が日本文字列と混在し、しかも全角文字と半角文字の英数文字列が混在するような場合にも、英数文字列が定ピッチであるかプロポーショナルピッチであるかを精度よく判別できる。

【００４３】定ピッチ英数文字列の文字矩形ピッチは文字によって大きく変動するため、この変動が反映されない一定の基準ピッチと文字矩形ピッチとの相対的な広狭を調べても、スペースを精度よく検出できないが、請求項６記載の発明によれば、注目した文字間の前と後の文字間の文字矩形ピッチの中の小さい方を、注目した文字間に対する基準ピッチとして用いるため、文字矩形ピッチの変動が大きい定ピッチ英数字文字列中のスペースを精度よく検出することができる。

【００４４】プロポーショナルピッチ英数文字列は文字矩形間隔が狭いため、ピッチ書式や文字種を考慮しない標準文字サイズと文字矩形間隔とを比較する方法では、スペース検出が難しかったが、請求項７記載の発明によれば、個々のプロポーショナルピッチ英数文字列毎に標準文字サイズを決定し、標準文字サイズに対する文字矩形間隔の比を閾値処理するため、プロポーショナルピッチの英数文字列中のスペースを精度よく検出できる。

【図面の簡単な説明】

【図１】本発明の一実施例の全体処理フローを示すフローチャートである。

【図２】本発明の一実施例のための装置構成の例を示すブロック図である。

【図３】図１中のスペース検出処理の概略を示すフローチャートである。

【図４】図３中のピッチ書式判定処理の内容を示すフローチャートである。

【図５】図３中のプロポーショナルピッチ文字列スペース検出処理の内容を示すフローチャートである。

【図６】図３中の定ピッチ文字列スペース検出処理の内容を示すフローチャートである。

【図７】スペース検出の具体例を説明するための図である。

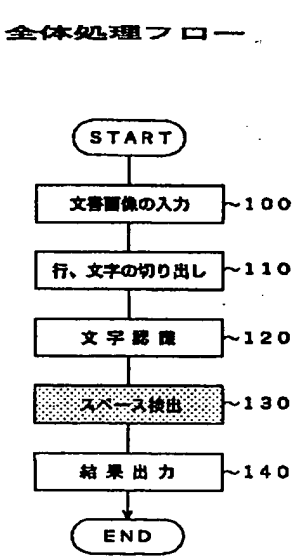
【図８】ピッチ書式判定のための標準文字サイズの決定方法の説明図である。

【図９】日英混在文書のスペース検出における課題を明らかにするための図である。

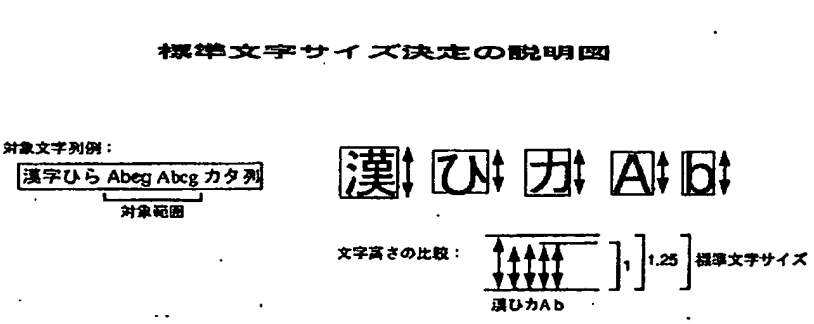
【符号の説明】

２００　画像入力部
２０２　行・文字切り出し部
２０３　文字認識部
２０４　対象範囲抽出部
２０５　ピッチ書式判定部
２０６　定ピッチ文字列スペース検出部
２０７　プロポーショナルピッチ文字列スペース検出部
２０８　画像メモリ
２０９　切り出しメモリ
２１０　文字辞書メモリ
２１１　結果メモリ
２１２　ワークメモリ
２１３　バス

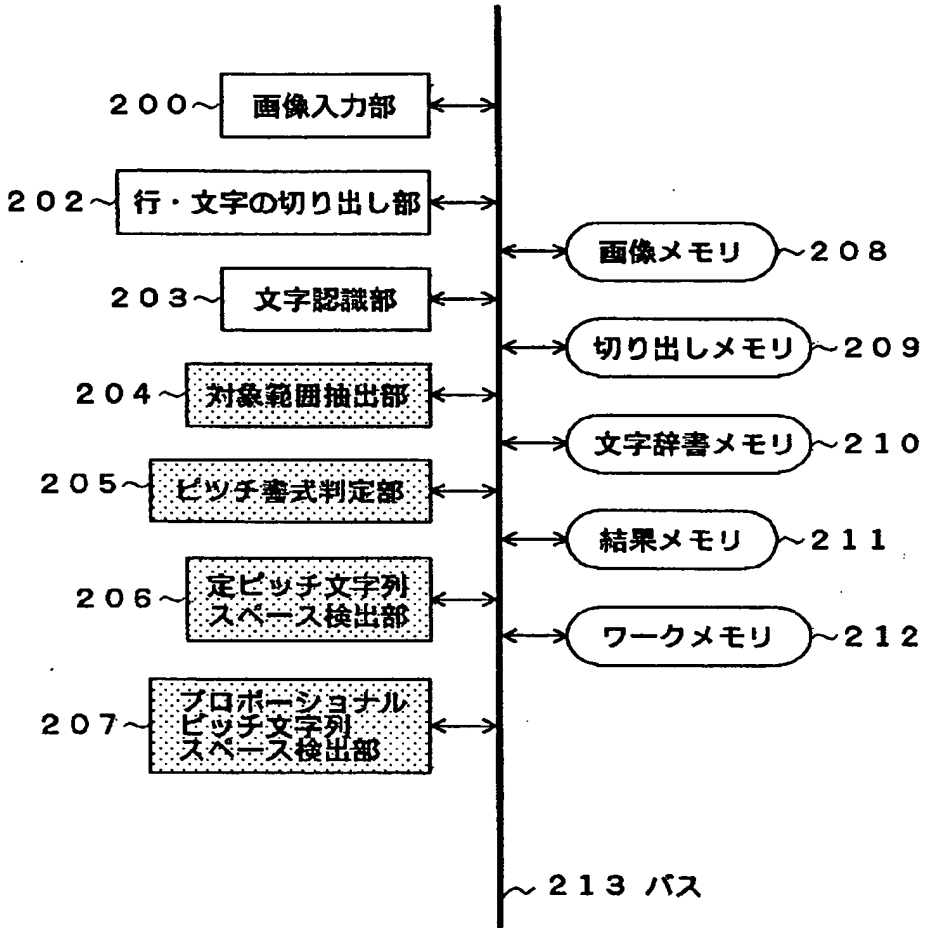
【図１】



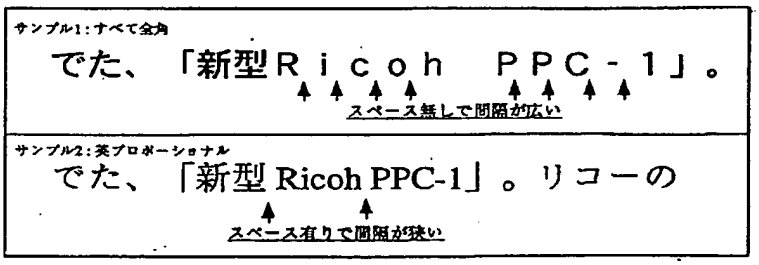
【図８】



【図2】

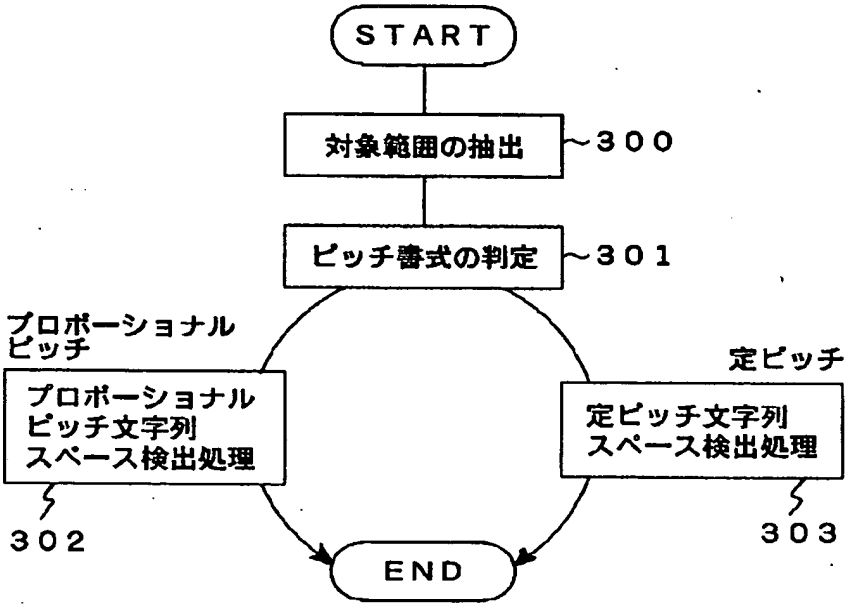

200
画像入力部
202
行・文字の切り出し部
203
文字認識部
204
対象範囲抽出部
205
ピッチ書式判定部
206
定ピッチ文字列スペース検出部
207
プロポーショナルピッチ文字列スペース検出部
画像メモリ
208
切り出しメモリ
209
文字辞書メモリ
210
結果メモリ
211
ワークメモリ
212
213 バス

【図9】


サンプル1：すべて全角
でた、「新型Ｒｉｃｏｈ　ＰＰＣ－１」。
スペース無しで間隔が広い
サンプル2：英プロポーショナル
でた、「新型 Ricoh PPC-1」。リコーの
スペース有りで間隔が狭い

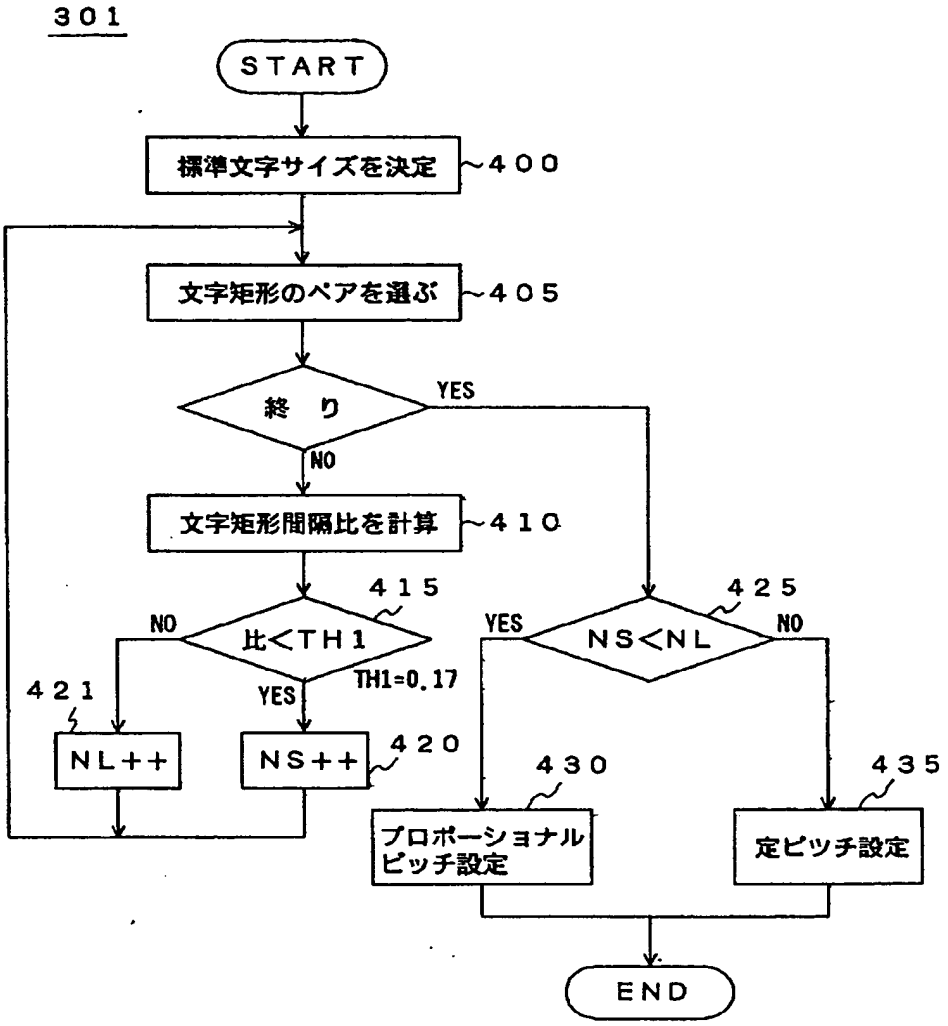
【図3】

スペース検出処理フロー



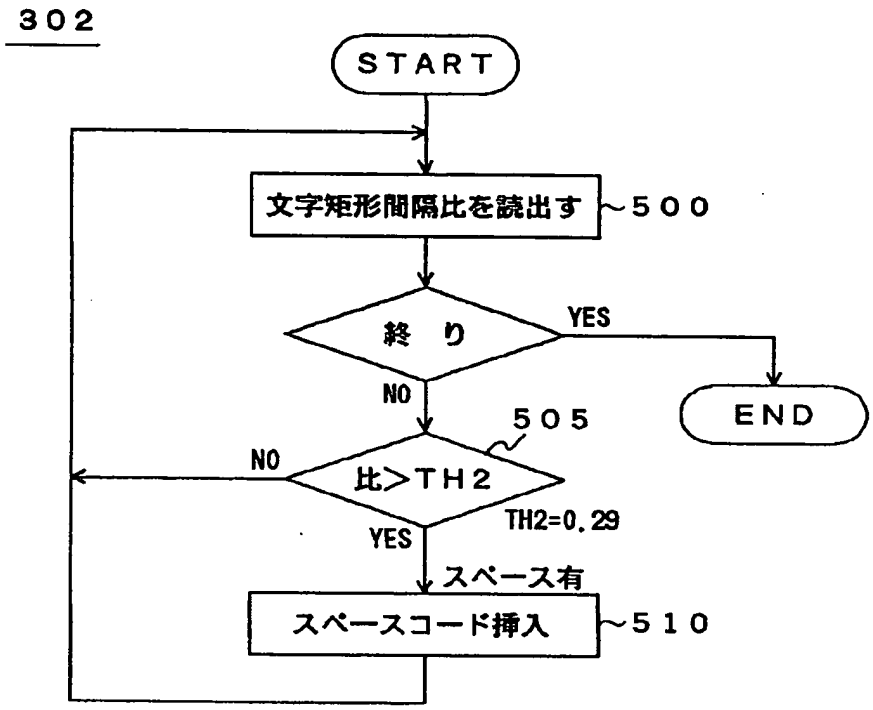
【図4】

ピッチ書式判定


301
START
標準文字サイズを決定
400
文字矩形のペアを選ぶ
405
終　り
YES
NO
文字矩形間隔比を計算
410
415
比<TH1
NO
YES
TH1=0.17
421
NL++
NS++
420
425
NS<NL
YES
NO
430
プロポーショナル
ピッチ設定
435
定ピッチ設定
END

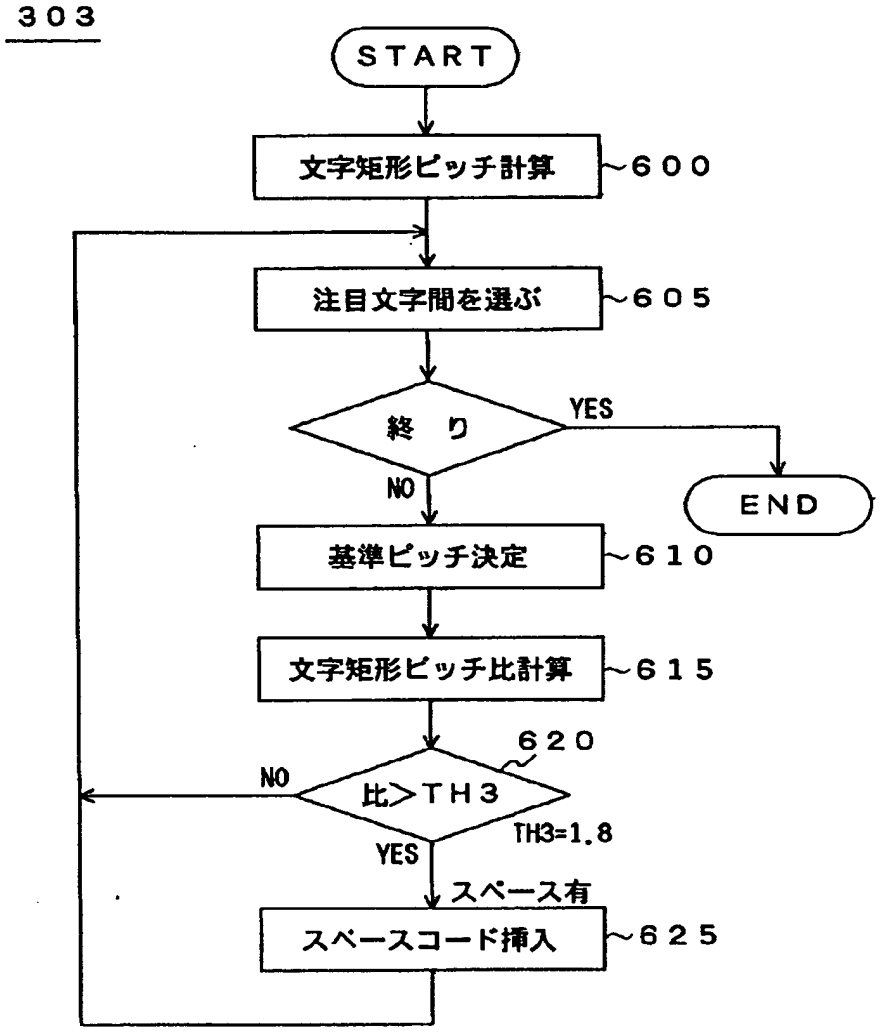
【図5】

プロポーショナルピッチのスペース検出



【図6】

定ピッチのスペース検出



【図7】

スペース検出の処理例